Team 17: Human Pose Prediction in the Wild

Mohammad Asim 7029636
Julian Gebhard 7027713
Michael Sonntag 2563456

1. Task and Motivation

Human pose prediction is about predicting a human's future pose based on its movement so far. One approach is using skeletal representations of the human body, which reduce the human body to a certain number of discrete points. Temporal sequences of these skeletal representations can be interpreted as movement and used for predicting future trajectory and pose. Anticipating humans poses and movement is crucial for robots when moving near humans. With knowledge about their future pose, a robot can avoid taking dangerous actions or movement which could harm the human, thus increasing the overall safety. By using state-of-the-art methods in deep learning and their ability to learn from a given data, we can shorten the reaction time of a robot by peeking into the near future for detecting any potential hazards. By making our approach adaptive to camera movements and considering the surroundings of the human, the approach could also be applied on cameras mounted on free moving robots/vehicles.

1.1. Related Work

One way of solving skeleton pose prediction and consider the surroundings of the human was done by T. Fujita and Y. Kawanishi in 2023. [1] They used given graph-convolutional networks (GCN) and extended them with so-called image-assisted attention (IAA). IAA uses image features extracted by a pre-trained network from an image of the surroundings corresponding to the last skeleton pose of the sequence to modify the skeleton features produced by the GCN.

Using transformers for pose prediction was performed by A. Martínez-González et. al. in 2021. [2] They modelled pose predictions as a sequence-to-sequence problem. By using a pre-generated query sequence in combination with a non-autoregressive Transformer they generate sequences of the future skeleton movement.

B. Parsaeifard et. al. [3] proposed decoupled representation for forecasting human pose from a sequence of input poses. Their method decouples the problem into local pose forecasting and global trajectory optimization using root relative representation for each pose. Primarily, they use LSTM for forecasting the global trajectory and generative approach with VAE for forecasting local poses. As such, their method then tries to optimize each jointly by minimizing L2-losses for the local relative poses and the root joint with an additional Kullback–Leibler divergence term to map the latent distribution to a Gaussian distribution for the local pose. They argue that with such decoupled representation, they can model the dependencies between the changes in the local poses and the global trajectories.

E. Vendrow et. al. [4] proposed a new paradigm for forecasting human poses in a multi-person scenarios using transformers. Their method uses joint-aware representation of human poses instead of traditional time-based sequences for multiple persons. With such representation, their method can perform attention on each joints and is able to distinguish between multiple humans.

The method presented by [1–4], is limited to fixed camera setup. This limits its applicability and versatility. Similarly, [2–4] does not consider the dynamics of the surrounding environment except for [1] which is still limited in a spatial setting and [1, 2, 4] do not consider decoupled representation of the human poses.

2. Goals

Our goal is to address these limitations by combining some of the key aspects mentioned in those papers in a more challenging setting.

- Considering the global spatial cues and extend it into a temporal setting to account for arbitrary camera setup and moving camera.
- Considering the local spatial cues and extend it to a temporal setting to account for moving objects or entities in the final prediction.
- Decouple the prediction problem into predicting global trajectory and local poses with the help of surrounding cues in a spatio-temporal setting.

We are hoping to develop a baseline model (using standard approaches) till the mid-term after preparing the input

data-loading, training and evaluation pipeline. By reusing some of the modules from the exercises, we can speed up the development and save time to focus on the method itself. We will evaluate the potential for this method and decide further improvements on the concept.

3. Methods

As an image feature extractor (local and global), we plan to use standard Vision Transformer (ViT) and for pose features (local and global) we have a couple of options to test from e.g. GCN-based or transformer-based feature extractors etc.

We divide the problem into two sub parts i.e., optimization of the root joint (global trajectory) and the trajectory of root-relative poses given a sequence of past poses and images. The intuition is to allow the global feature of the pose to attend to the spatio-temporal encoding of itself and the global/local features of the image sequence such that they are able to adapt to the global changes e.g. movement of the camera and to relate to the local surroundings and itself e.g. movement of the human.

The local features of the poses are more likely to be dependent on the depth (via scale) and the local surroundings e.g., chairs, foot-paths, benches, entities etc. and the spatio-temporal encoding of itself.

This is where our approach differs from the previous work using the idea from [3] for decoupling the problem into local and global approach, to [1] for leveraging the surrounding cues and extending it to dynamic scenes in a spatio-temporal setting. As such, we think this dual conditioning will help to enhance the predictions in a meaningful way. Figure 1 represent the general idea.

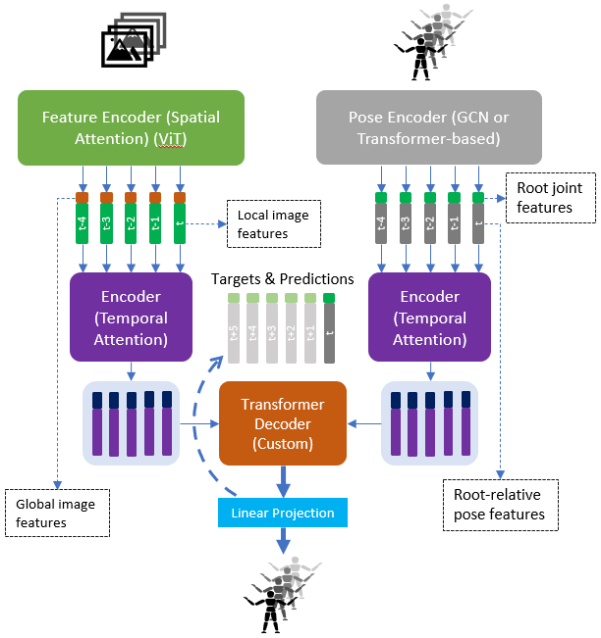

Figure 1. Initial abstract representation of the proposed approach.

4. Datasets

For training and proving our approach of wild pose future prediction model a specific type of time-series data is needed. A sequence of image displaying humans taking common poses. As this work is focusing on the elaboration of a machine learning model to detect patterns in the underlying distribution we choose the benchmark to be the open source dataset "3D Poses in the Wild" captured by T. Von Marcard et al. [5]. It has been used to recover 3D human poses. The "dataset consisting of more than 51, 000 frames with accurate 3D pose in challenging sequences, including walking in the city, going up-stairs, having coffee or taking the bus" [5]. This variate of different sequences will be the benchmark for the developed model.

5. Evaluation

There is no goal to apply the model in real world applications, so the evaluation will be only capture the statistical results that can be achieved on the open source dataset. The performance of the model will be evaluated by comparing different error measurements. Many standard loss functions give an insight into the performance, e.g. MSE, MAE. In this case the Mean Joint Position Error (MPJPE) have been used in related problems. [4, 5]

$$\text{MPJPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{J}\sum_{j=1}^{J}\|\mathbf{P}_i^j - \mathbf{G}_i^j\|$$

The method is based on the Euclidean distance. Here P is the predicted position, G is the ground truth, J is the total number of joins and N is the number of samples. Afterwards a comparison to other methods that have been applied to related benchmarks, e.g. posing with static camera, will be performed.

Moreover, it is not uncommon to create new metrics by combining existing ones to evaluate research gaps in a more dynamic way. Especially because we are using two models, which leads more parameters to analyze.

References

[1] Tomohiro Fujita and Yasutomo Kawanishi. Future pose prediction from 3d human skeleton sequence with surrounding situation. *Sensors*, 23(2), 2023. 1, 2

[2] Ángel Martínez-González, Michael Villamizar, and Jean-Marc Odobez. Pose transformers (POTR): human motion prediction with non-autoregressive transformers. *CoRR*, abs/2109.07531, 2021. 1

[3] Behnam Parsaeifard, Saeed Saadatnejad, Yuejiang Liu, Taylor Mordan, and Alexandre Alahi. Learning decoupled rep-

resentations for human pose forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2294–2303, October 2021. 1, 2

[4] Edward Vendrow, Satyajit Kumar, Ehsan Adeli, and Hamid Rezatofighi. Somoformer: Multi-person pose forecasting with transformers, 2022. 1, 2

[5] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, pages 601–617, 2018. 2